Gathering situated dialogue in the field

ANDREW GARGETT[1] & SAM HELLMUTH[2]
School of Computer Science, University of Birmingham[1] & Dept of Language & Linguistic Science, University of York[2]

1. BACKGROUND

Documentation of communicative behaviour across languages seems at a crossroads. While methods for collecting data on spoken or written communication, backed up by computational techniques, are evolving, the actual data being collected remain largely the same. Recently, this general trend has been defied by some innovative researchers, who are often investigating language in the field (e.g. see various papers collected in Enfield & Stivers 2007). Inspired by such efforts, we report here our attempt to collect data for situated linguistic interaction, employing a portable format designed to increase range and flexibility of doing such collections in the field. Our motivation is to combine this with a parallel data set for a typologically distinct language, in order to contribute a parallel corpus of situated language use.

Our motivation for developing a corpus directly incorporating such information comes from our observation that established corpora of instruction-giving dialogues (e.g. TRAINS[3], Map Task[4]) typically require interlocutors to interact in a relatively stable yet static situation, with very little engagement in or monitoring of the surrounding environment. Communication as characterised by such collections is relatively less situated, interactive in varying degrees, and typically yields a limited range of data (e.g. typically verbal code and vocal channel, but also at times including paralinguistic, as well as non-vocal gestures, gaze, proxemics, etc). However, there is rarely much detail about the actions taking place in and around language, which has led to a paucity of information about how people use situational features, including environment and actions, during such forms of dialogue. A direct consequence of this is that our understanding of how interlocutors use such information is quite impoverished. Extending models of interaction to incorporate such information is likely to provide qualitatively distinct accounts of what is going on in dialogue. This is the motivation for the approach to collecting such data which is presented in this paper.

By way of providing sufficient background to our approach, additional conceptual details are required. By dialogue, we mean interactive and co-

[1] Email: A.D.Gargett@cs.bham.ac.uk
[2] Email: Sam.Hellmuth@york.ac.uk
[3] http://www.cs.rochester.edu/research/speech/trains.html
[4] http://groups.inf.ed.ac.uk/maptask/

ordinated dyadic conversation. More details of the dialogue making up our corpus are given below, but essentially they involve people engaged in giving and following instructions. These dialogues are situated in that information about both environment and actions taken by interlocutors is recoverable.[5]

2. METHOD

This paper reports on the progress of a novel corpus of human-to-human real-time spoken instruction giving in 3D environments for Gulf Arabic: Dialogue in Virtual Environments (DiVE-Arabic).[6] The 3D worlds used to record this corpus were the same as those used by Stoia et al. (2008), for the SCARE corpus,[7] an English corpus of instruction giving in a virtual world. The SCARE corpus is a collection of instruction giving dialogues in a virtual world made up of two levels, each with between 7 and 9 rooms, and these rooms having buttons for opening cabinets that contained objects to be retrieved (see Example (1) below for a screenshot). The corpus employs the QuakeII gaming software. The corpus consists of 15 sessions, with interlocutors taking roles of either instruction giver (IG) or instruction follower (IF). They had to complete a series of 5 simple tasks (retrieving or manipulating objects), with the IG verbally guiding the IF through the world, but only the IG having access to a map of the world, and a list of the tasks to be completed. The 19 male and 11 female participants had an average age of 30, and identified as native speakers of North American English. Sessions ranged from 10 minutes in length to over half an hour.

Our corpus collection was carried out alongside other data gathering for the Intonational Variation in Arabic (IVAr) project[8] (Hellmuth 2014), itself inspired by earlier work on English,[9] and served the purpose of collecting situated dialogue in Gulf Arabic (one of the dialects covered by IVAr). In Al Ain (UAE) and Buraimi (Oman), we recorded dialogue between participants trying to solve instruction giving tasks within a virtual world. For this corpus collection, we replicated the SCARE task and number of participants recorded, while also approximating as closely as possible the population demographics. For the Gulf Arabic corpus, sessions ranged from just under 10 minutes to just over half an hour. The 20 female and 10 male participants were university students aged in their early to late-twenties, who interacted in same gender pairs. The DiVE-Arabic corpus is then commensurate in terms of extra-linguistic factors with the SCARE corpus (except that it involves no gender mixing of interlocutors). A key

[5] We would like to acknowledge the help of Dr Rana Alhussein Almbark (University of York) in preparing materials used during the data collection. We would also like to thank the Department of Linguistics, University of the UAE, Al Ain, for hosting S.H. during her visit to carry out fieldwork.

[6] http://www.york.ac.uk/res/iva/dive

[7] http://slate.cse.ohio-state.edu/quake-corpora/scare/

[8] http:// www.york.ac.uk/res/ivar

[9] http:// www.phon.ox.ac.uk/IViE/

motivation for replication was to make possible an eventual parallel English-Arabic corpus of situated dialogue when the SCARE and DiVE corpora are combined, making this a highly unique contribution to the area of cross-linguistic dialogue studies.

We recorded audio signals of each interlocutor, screen capture of video of what each participant sees during their interaction, as well as detailed information about instruction follower movements in the virtual world (the computer continuously records orientations and positions in the virtual world). In more detail, this recorded information includes:

a. Spoken Arabic, one channel per speaker (using a Marantz PMD661 solid state digital audio recorder, and two Shure SM10 unidirectional head-mounted microphones). Annotation is currently in progress, and is aligned with the signal.
b. Actions of the instruction follower in a virtual world (the same world as used for the SCARE corpus). Instruction giving sessions took place through the Jake2[10] platform (a freely available Java version of QuakeII), and these sessions were recorded on a PC laptop running Windows 7. All actions will be automatically acquired from the computer log files, and incorporated in the corpus.
c. Video signal of the monitor output for the virtual world, showing the location and actions of the instruction follower as they move through the world, and both participants can view this while interacting. There are no immediate plans to annotate the video data, although this data will also be incorporated in the corpus.

In summary, all of this information captures exactly what the interlocutors said to each other while interacting, exactly what they could see, and exactly what they did.

3. RESULTS

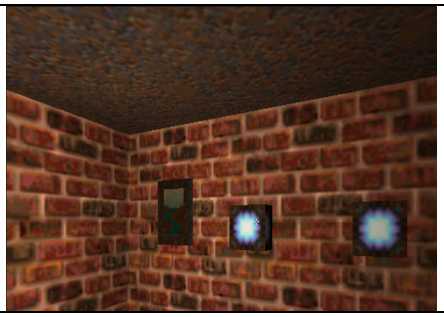
Richness in the resulting data comes from interlocutors talking about assigned tasks, while the instruction follower moves around the virtual environment to solve problems that are encountered. Instruction givers know things which instruction followers don't (e.g. only instruction givers have a map of the world), *but only instruction followers can move*, forcing interlocutors to work together. We look at two case studies: the first considers how interlocutors can monitor actions in order to disambiguate prosody; the second examines how interlocutors employ prosodic features to distinguish between questions and requests.

[10] http://jake2.cvs.sourceforge.net/

3.1 Case Study 1

Figure 1 below illustrates how situated action can disambiguate prosody. Here the instruction follower M1 is trying to complete an assigned task of moving the picture on the wall, and only one button performs this action; the correct button is marked on the instruction-giver's map.

Figure 1
Example illustrating Case Study 1 (including picture of items referred to)



1.	IGM2:	ʔajwa iðˤɣɑtˤ ʕala:: *yes push on…*
2	IFM1:	ʔajji waħdˤɑ (0.4) [il-ʔu:la] *which one? the-first?*
3	IGM2:	[az-] (0.3) izzur- izzur iθθa:ni *the- the-button the-button the-second*
	⇒	(1.3)
4	IGM2:	ʔajwa ha:ða laʔ (1.4) *yes this no*
	⇒	(1.3)
5	IGM2:	ʃu:f ajwa sˤɑħ *look yes correct*
6	IFM1:	nze:n Ok

Multi-level recording provides a record of participant actions as well as utterances, and the links between these. This unlocks vital information: it would be impossible to interpret the interactional value of the sequences produced by M2 in lines 4 and 5 in Figure 1 above, based on audio alone, but with the information provided by the video/movements record, we can make inferences about the interaction involved. In a conventional approach presenting textual modality alone, the two major silences, preceding lines 4 and 5, respectively, could only really be interpreted via the verbal code; in both cases, we have a lengthy within speaker pause and no apparent response from the interlocutor. This is potentially a sign of trouble, but is this so in this case?

Our approach provides added situational information, revealing an additional dimension of such silences, which can disambiguate. During a linguistic silence, the instruction follower may either move around, push buttons, turn and change orientation, or indeed stay absolutely still and do nothing; that is, linguistic silences can now be interpreted via an additional dimension, in terms of the non-linguistic actions. For example, the instruction follower saying nothing but moving is likely to be interpreted quite differently by an Instruction Giver (IG) monitoring the IF's behaviour, compared to wen the IF says and does nothing. In the present example, the IF is active during both silences: during the first silence IF changes orientation several times with gaze towards different buttons; during the second silence there is backwards movement away from the button and a change of orientation towards the picture (to see if the task has been completed).

In a prior study of the SCARE corpus, Gargett (2012) showed that, in English, on average, an instruction giver is more likely to produce a further verbal response after a linguistic silence if there is no accompanying action, than if there is accompanying action. Full elaboration of the DiVE-Arabic corpus will allow us to determine whether the same pattern is in fact also observed in Arabic, or whether there are cross-linguistic differences in such matters.

3.2 Case Study 2

The case study involves the way in which dialogue context cues can yield independent evidence to support analysis of prosodic phenomena. Figure 2 illustrates this patterning of context and prosody.

Prosodic analysis shows that the speakers systematically realise the last accented syllable of utterances at different levels of pitch, finishing at a pitch level which is either high (H), mid (M) or low (L) in their pitch range. Figure 3 below illustrates the difference in final pitch, M vs H, for speaker IGM2.

Figure 2
Example illustrating Case Study 2 (including picture of items referred to)

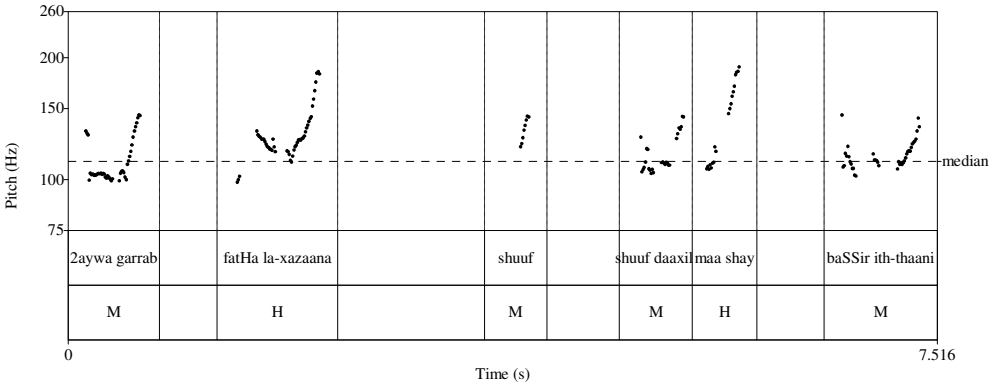
1	IFM1:	ʔaðˤɣɑtˤ ʕalay-h haːða (0.7) *push on-it this?*	H
2	IGM2:	ʔaywa garrab (0.8) *yes try*	M
3	IGM2:	fatħa l-xazaːna (0.7) *opened the-cupboard?*	H
4	IFM1:	ʔaywa (0.6) *Yes*	L
5	IGM2:	ʃuːf (0.8) *look*	M
6	IGM2:	ʃuːf daːxil *look inside*	M
7	IGM2:	maː ʃay (0.6) *NEG thing?*	H
8	IFM1:	maː fiː-ʃ *NEG there-NEG*	L
9	IGM2:	basˤsˤir ith-thaːni (2.3) *look the-second*	M
	⇒		
10	IGM2:	maːshi ʔaywa ʃiːl haːða *ok yes carry this*	L

Inspection of the verbal text suggests that there is a correlation between choice of final pitch level of a contribution to the dialogue and its function or role; responses are low (L), requests are mid (M) and questions are high (H). In our corpus we are able to provide an additional layer of evidence to support this classification. The video/movements data shows that contributions with H and M final pitch are treated differently by the interlocutor: H-final contributions (putative questions) always receive a verbal response (generally unaccompanied by an action response), but M-final contributions (putative requests) always precede actions (and generally no verbal response). One explanation for this kind of systematic difference is that interlocutors are using prosody to distinguish

questions, i.e. requests for information, from requests for action. This difference is revealed only because we can recover both linguistic and non-linguistic information from our corpus.

Figure 3

Pitch trace for lines 2-9 of case study 2, for speaker IGM2 only. Axes show this speaker's min/max/median pitch range measured in whole conversation.



3.3 Summary of results

Our corpus makes several novel contributions, which have arisen in order that we may address specific research questions:

a. As part of the IVAr project, we designed our data collection to capture information about how intonation links to other linguistic levels of situated dialogue in Arabic. Note that SCARE focused on referring expressions in English.
b. A completely new dimension of our corpus is that it enables exploration of how intonation links to the actions being undertaken in the situation in which the dialogue takes place.
c. Additionally, unlike comparable approaches (e.g. GIVE-2[11]), we are able to explore in detail the interaction between all levels of communicative behaviour, including spoken language, as well as actions carried out while interlocutors are interacting, enabling richer possibilities for investigating how such behaviour is grounded in the surrounding environment during communication.

However, it should be stressed that augmenting the SCARE corpus like this in no way impedes the compatibility of the two corpora. Indeed, we have also analysed the original audio signal from SCARE, splitting this into two channels, whereby we are able to incorporate tracks for both instruction-giver and follower,

[11] http://www.give-challenge.org/research/page.php?id=give-2-corpus

into the resulting database we have built for the combined corpus (after checking and cleaning this data). [12]

Our plan is to construct the combined corpus as a stand-off database using the Nite NXT toolkit[13] (Carletta et al.). The Nite NXT approach is particularly useful for us due to its rich structuring of data, including a data set model for structuring a corpus in terms of (i) observations, (ii) agents, (iii) the interaction, as well as (iv) annotations of the signal. We further divide the signal into segmental and supra-segmental components, each being stored separately in line with the stand-off approach. In particular, the observations can be multi-layered, either directly aligned to the timing level, or else symbolically linked to other levels (e.g. annotations of dialogue acts can be linked to actual utterances, which in turn can be directly aligned with the timing of the original audio and video signal). Aside from allowing us to adequately model the rich information from the dialogue data, this also allowed access to a very useful library of Java classes bundled with the Toolkit (e.g. for searching NXT-formatted corpus files).

Using a range of technology, we have built tools for collecting manual annotations, as well as automatically collecting analyses (such as automatic alignment), into a comprehensive XML database, that can link recorded data, annotations of this data, as well as other relevant information, in a multi-dimensional way. As shown in Gargett (2012), this way of presenting the information enables a flexible set-up for carrying out analyses across modalities.

4. CONCLUSION

Our project has the potential to allow researchers to recover both linguistic and non-linguistic information about the situatedness of dialogue, which has previously been unavailable. We aim to provide a genuinely cross-linguistically parallel corpus, whereby results for one language can be usefully compared to those in the other.

One slight limitation for the current setup which we are now addressing is that the task itself, while well-suited to collecting data sufficient for modelling the range of phenomena we are interested in, may not result in amounts of data sufficient for all possible dialogue projects. We are currently fully revising the task and setup of the worlds, in order to tackle this issue, with the aim of trialling it on an even greater range of languages (e.g. Mandarin and Tamil, as well as English and Arabic). A further limitation is that there may be an issue to do with mobility in very remote research locations, and we have plans to develop a

[12] This was carried out in the course of background research for Gargett (2012).

[13] http://groups.inf.ed.ac.uk/nxt/

version of this approach that may be deployed on mobile technology (e.g. tablets, smart phones).

REFERENCES

Carletta, J., S. Evert, J. Kilgour, C. Nicol, D. Reidsma, J. Robertson & H. Voormann, Documentation for the NITE XML Toolkit. http://http://groups.inf.ed.ac.uk/nxt/documentation.shtml.

Enfield, N. J. & Tanya Stivers. 2007. *Person Reference in Interaction: Linguistic, Cultural and Social Perspectives*. Cambridge: Cambridge University Press.

Gargett, Andrew. 2012. Feedback and activity in dialogue: signals or symptoms? *Proceedings of the Interdisciplinary Workshop on Feedback Behaviors in Dialog*, September 7 -8, 2012, Oregon.

Hellmuth, Sam. 2014. Dialectal Variation in Arabic Intonation: Motivations for a Multi-Level Corpus Approach. In: Farwaneh, Samira and Hamid Ouali (eds.), *Perspectives on Arabic Linguistics XXIV-XXV*, 63–90. Amsterdam: John Benjamins.

Stoia, Laura, Darla M. Shockley, Donna K. Byron, & Eric Fosler-Lussier. 2008. SCARE: A Situated Corpus with Annotated Referring Expressions. *Proceedings of the 6th International Conference on Language Resources and Evaluation*.